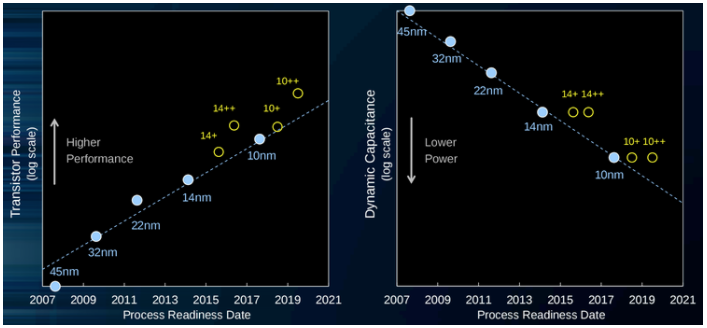
Future trends

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Codename** | **Broadwell** | Skylake | Kabylake | Coffeelake | Cannonlake | Icelake | Tigerlake | Sapphire Rapids |  |
| **Arch type** | CPU | CPU | CPU | CPU | CPU | CPU | CPU | CPU | CPU |
| **Designer** | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel |
| **Manufacturer** | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel |
| **Introduction** | 2014 | 2015 | 2017 | 2017 | 2017 | 2018 | 2019 | 2020 | 2020 |
| **Process** | 14nm | 14nm | 14nm | 14nm | 10nm | 10nm | 10nm | 10nm | 7nm |
| **Predecessor** | Haswell | Broadwell | Skylake | Kabylake | Kabylake | Cannonlake | Icelake | Tigerlake | Saphire Rapids |
| **Succesor** | Skylake | Kabylake | Coffeelake, Cannon lake | Icelake | Icelake | Tigerlake | Sapphire Rapids | |  |
| **Cycle** | **Process** | **Architecture** | Optimization | Optimization | Process | Architecture | Optimization | Optimization | Process |

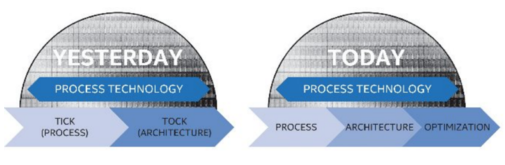


Process-Architecture-Optimization (PAO) - Intel

**Process-Architecture-Optimization** is a development model introduced by [Intel](https://en.wikichip.org/wiki/Intel) for their mainstream microprocessors in [2016](https://en.wikichip.org/wiki/2016) following the phase-out of their [Tick-Tock](https://en.wikichip.org/wiki/intel/tick-tock) model. The change is a result of the increase in cost and complexity of advancing lithography processes in the past decade. Under the new model the amount of time utilized for any given process technology is lengthened as [Moore's Law](https://en.wikichip.org/wiki/Moore%27s_Law) increases in complexity with smaller [nodes](https://en.wikichip.org/wiki/technology_node).

Under the Process-Architecture-Optimization Model:

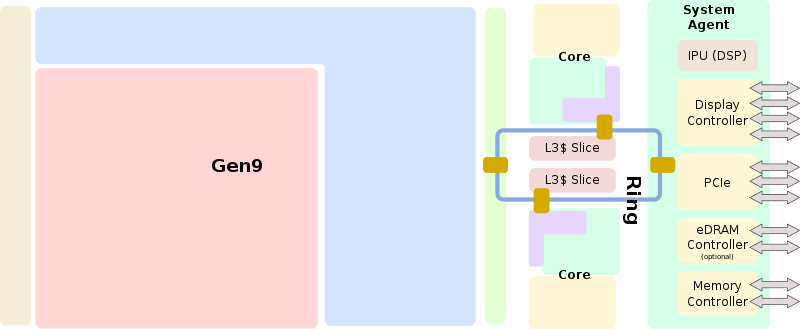
* **Process** - With each process, Intel advances their manufacturing [process technology](https://en.wikichip.org/wiki/process_technology) in line with [Moore's Law](https://en.wikichip.org/wiki/Moore%27s_Law). Each new process introduces higher transistor density and a generally a plethora of other advantages such as higher performance and lower power consumption. During a "process", Intel retrofits their [previous](https://en.wikichip.org/wiki/intel/microarchitectures" \o "intel/microarchitectures)[microarchitecture](https://en.wikichip.org/wiki/microarchitecture) to the new process which inherently yields better performance and energy saving. During a "process", usually just a few features and improvements are introduced.
* **Architecture** - With each architecture, Intel uses the their latest manufacturing [process technology](https://en.wikichip.org/wiki/process_technology) from their "process" to manufacture a newly designed [microarchitecture](https://en.wikichip.org/wiki/microarchitecture). The new microarchitecture is designed with the new process in mind and typically introduces Intel's newest big features and functionalities. New [instructions](https://en.wikichip.org/w/index.php?title=instruction_set&action=edit&redlink=1) are often added during this cycle stage.
* **Optimization** - With each optimization, Intel improves upon their [previous](https://en.wikichip.org/wiki/intel/microarchitectures) microarchitecture by introducing incremental improvements and enhancements without introducing any large charges. Additionally the process itself enjoys various refinements as it matures. (For example with [Kaby Lake](https://en.wikichip.org/wiki/intel/microarchitectures/kaby_lake" \o "intel/microarchitectures/kaby lake), an optimized process called "14 nm+" is used. The enhanced process had a number of transistor-level modifications done to it (e.g. taller fins) allowing for higher frequency at identical voltage levels.)



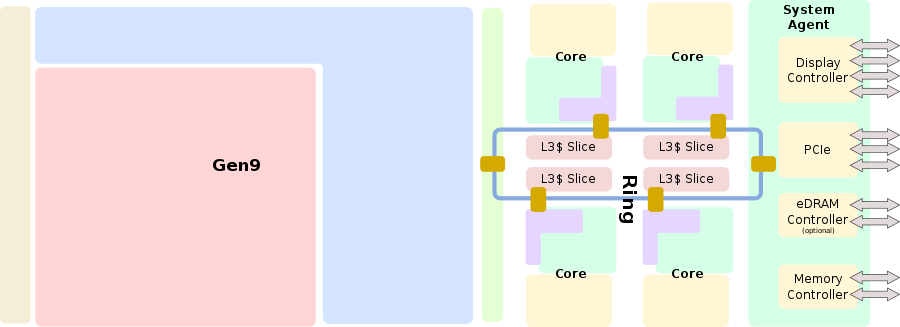


|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Codename | Haswell | Broadwell | Skylake | Kabylake | Coffeelake | Cannonlake | Icelake | Tigerlake | Sapphire Rapids |  |
| Arch type | CPU | CPU | CPU | CPU | CPU | CPU | CPU | CPU | CPU | CPU |
| Designer | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel |
| Manufacturer | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel | Intel |
| Introduction | 2013 | 2014 | 2015 | 2017 | 2017 | 2017 | 2018 | 2019 | 2020 | 2020 |
| Process | 22nm | 14nm | 14nm | 14nm | 14nm | 10nm | 10nm | 10nm | 10nm | 7nm |
| Predecessor | **Ivy Bridge** | Haswell | Broadwell | Skylake | Kabylake | Kabylake | Cannonlake | Icelake | Tigerlake | Saphire Rapids |
| Succesor | **Broadwell** | Skylake | Kabylake | Coffeelake, Cannon lake | Icelake | Icelake | Tigerlake | Sapphire Rapids | |  |
| Cycle | **Tock** | **Process** | **Architecture** | Optimization | Optimization | Process | Architecture | Optimization | Optimization | Process |

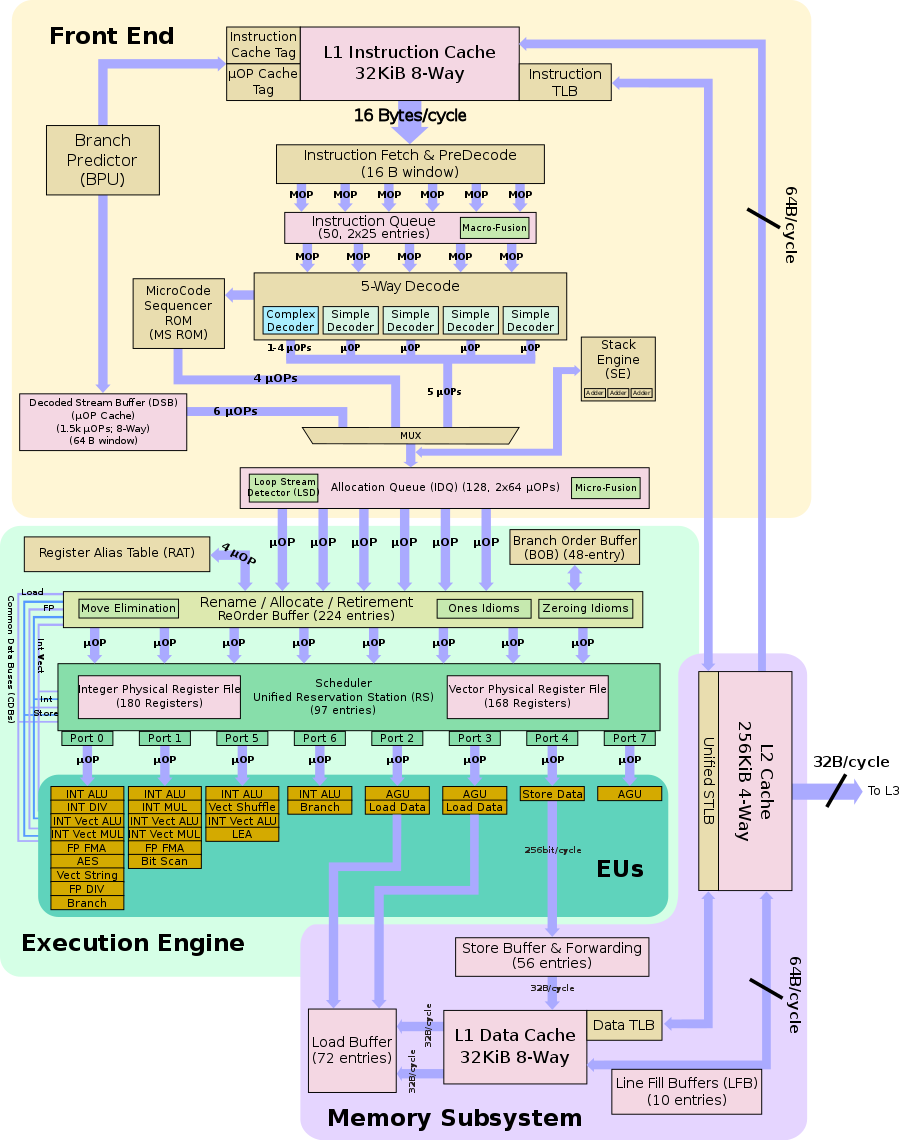
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Codename | Haswell | Broadwell | Skylake | Kabylake | Coffeelake |
| Arch type | CPU | CPU | CPU | CPU | CPU |
| Designer | Intel | Intel | Intel | Intel | Intel |
| Manufacturer | Intel | Intel | Intel | Intel | Intel |
| Introduction | 2013 | 2014 | 2015 | 2017 | 2017 |
| Process | 22nm | 14nm | 14nm | 14nm | 14nm |
| Predecessor | **Ivy Bridge** | Haswell | Broadwell | Skylake | Kabylake |
| Succesor | **Broadwell** | Skylake | Kabylake | Coffeelake, Cannon lake | Icelake |
| Cycle | **Tock** | **Process** | **Architecture** | Optimization | Optimization |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| L0 MicroOp Cache |  |  | **1,536 µOPs** | **1,536 µOPs** | **1,536 µOPs** |
|  | Associativity |  | 8-way set associative | 8-way set associative | 8-way set associative |
|  | Sets |  | **32** | **32** | **32** |
|  | Block size |  | 6-µOP | 6-µOP | 6-µOP |
|  |  |  | statically divided between threads, per core, inclusive with L1I | statically divided between threads, per core, inclusive with L1I | statically divided between threads, per core, inclusive with L1I |
|  |  |  |  |  |  |
| L1 Instruction Cache | **Cache Size** |  | **32KiB/core** | **32KiB/core** | **32KiB/core** |
|  | **Associativity** |  | 8-way set associative | 8-way set associative | 8-way set associative |
|  | **Sets** |  | **64** | **64** | **64** |
|  | **Block size** |  | **64B** | **64B** | **64B** |
|  |  |  | **Shared 2Thread/core** | **Shared 2Thread/core** | **Shared 2Thread/core** |
|  |  |  |  |  |  |
| L1 Data Cache | **Cache Size** |  | **32KiB/core** | **32KiB/core** | **32KiB/core** |
|  | **Associativity** |  | **8-way** | **8-way** | **8-way** |
|  | **Sets** |  | **64** | **64** | **64** |
|  | **Block size** |  | **64B** | **64B** | **64B** |
|  |  |  | **Shared 2Thread/core** | **Shared 2Thread/core** | **Shared 2Thread/core** |
|  | **Fastest load-to-use** |  | **4 cycles** | **4 cycles** | **4 cycles** |
|  | **Complex addresses** |  | **5 cycles** | **5 cycles** | **5 cycles** |
|  | **Load Bandwidth** |  | **64B/cycle** | **64B/cycle** | **64B/cycle** |
|  | **Store Bandwidth** |  | **32B/cycle** | **32B/cycle** | **32B/cycle** |
|  | **Update policy** |  | **Write back** | **Write back** | **Write back** |
|  |  |  |  |  |  |
| L2 Unified | **Cache Size** |  | **256KiB/core** | **256KiB/core** | **256KiB/core** |
|  | **Associativity** |  | **4-way set associative** | **4-way set associative** | **4-way set associative** |
|  | **Sets** |  | **1024** | **1024** | **1024** |
|  | **Block size** |  | **64B** | **64B** | **64B** |
|  |  |  | **Non-Inclusive** | **Non-Inclusive** | **Non-Inclusive** |
|  | **Fastest load-to-use** |  | **12 cycles** | **12 cycles** | **12 cycles** |
|  | **Load Bandwidth** |  | **64B/cycle** | **64B/cycle** | **64B/cycle** |
|  | **Update policy** |  | **Write back** | **Write back** | **Write back** |
|  |  |  |  |  |  |
| L3 Unified | **Cache Size** |  | **2MiB/core** | **2MiB/core** | **2MiB/core** |
|  |  |  | **shared across all cores** | **shared across all cores** | **shared across all cores** |
|  | **Associativity** |  | 16-way set associative | 16-way set associative | 16-way set associative |
|  | **Block size** |  | **64B** | **64B** | **64B** |
|  |  |  | **Inclusive** | **Inclusive** | **Inclusive** |
|  | **Fastest load-to-use** |  | **42 cycles** | **42 cycles** | **42 cycles** |
|  | **Load Bandwidth** |  | **32B/cycle** | **32B/cycle** | **32B/cycle** |
|  | **Store Bandwidth** |  | **32B/cycle** | **32B/cycle** | **32B/cycle** |
|  | **Update policy** |  | **Write back** | **Write back** | **Write back** |
|  |  |  |  |  |  |
| Side Cache | **Cache Size** |  | **64MiB/core** | **64MiB/core** | **64MiB/core** |
|  | **eDRAM** |  | **128MiB** | **128MiB** | **128MiB** |
|  |  |  | **per package** | **per package** | **per package** |
|  |  |  | **only Iris Pro GPU** | **only Iris Pro GPU** | **only Iris Pro GPU** |
|  | **Load Bandwidth** |  | **32B/cycle** | **32B/cycle** | **32B/cycle** |
|  | **Store Bandwidth** |  | **32B/cycle** | **32B/cycle** | **32B/cycle** |
|  | **Update policy** |  | **Write back** | **Write back** | **Write back** |
|  |  |  |  |  |  |
| System DRAM | **No of Channel** |  | **2** | **2** | **2** |
|  | **size** |  | **8 B/cycle/channel** | **8 B/cycle/channel** | **8 B/cycle/channel** |
|  |  |  | 42 cycles + 51 ns latency | 42 cycles + 51 ns latency | 42 cycles + 51 ns latency |
| TLB |  |  |  |  |  |
| L1 ITLB | 4 KiB page |  | 128 entries | 128 entries | 128 entries |
|  |  |  | 8-way set associative | 8-way set associative | 8-way set associative |
|  |  |  | dynamic partitioning | dynamic partitioning | dynamic partitioning |
|  | 2 MiB / 4 MiB page |  | 8 entries per thread | 8 entries per thread | 8 entries per thread |
|  |  |  | fully associative | fully associative | fully associative |
|  |  |  | Duplicated for each thread | Duplicated for each thread | Duplicated for each thread |
|  |  |  |  |  |  |
| L1 DTLB | 4 KiB page |  | 64 entries | 64 entries | 64 entries |
|  |  |  | **4-way set associative** | **4-way set associative** | **4-way set associative** |
|  |  |  | fixed partition | fixed partition | fixed partition |
|  | 2 MiB / 4 MiB page |  | 32 entries | 32 entries | 32 entries |
|  |  |  | 4-way set associative | 4-way set associative | 4-way set associative |
|  |  |  | fixed partition | fixed partition | fixed partition |
|  | 1GiB page |  | 4 entries | 4 entries | 4 entries |
|  |  |  | fully associative | fully associative | fully associative |
|  |  |  | fixed partition | fixed partition | fixed partition |
|  |  |  |  |  |  |
| L2 TLB | 4 KiB + 2 MiB page |  | 1536 entries | 1536 entries | 1536 entries |
|  |  |  | 12-way set associative | 12-way set associative | 12-way set associative |
|  |  |  | fixed partition | fixed partition | fixed partition |
|  | 1 GiB page |  | 16 entries | 16 entries | 16 entries |
|  |  |  | 4-way set associative | 4-way set associative | 4-way set associative |
|  |  |  | fixed partition | fixed partition | fixed partition |



**Entire SoC Overview (dual)**



#### Entire SoC Overview (quad)

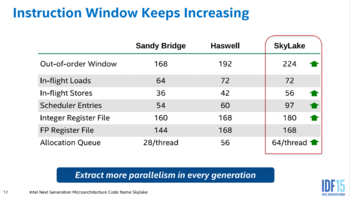


#### Individual Core

## **Architecture[[edit](https://en.wikichip.org/w/index.php?title=intel/microarchitectures/skylake_(client)&action=edit&section=8" \o "Edit section: Architecture)]**

Overall Skylake builds upon Intel's previous microarchitecture, [Broadwell](https://en.wikichip.org/wiki/intel/microarchitectures/broadwell), but includes a wider and more beefed up front end, more optimized execution engine, and numerous other enhancements. Intel designed Skylake to encompass a wide range of devices and applications with a large emphasis on mobile with models ranging from as low as 4.5 W to as high as 100 W.

### Key changes from [Broadwell](https://en.wikichip.org/wiki/intel/microarchitectures/broadwell)



### Key changes from [Broadwell](https://en.wikichip.org/wiki/intel/microarchitectures/broadwell)**[**[**edit**](https://en.wikichip.org/w/index.php?title=intel/microarchitectures/skylake_(client)&action=edit&section=9)**]**

* 8x performance/watt over [Nehalem](https://en.wikichip.org/wiki/intel/microarchitectures/nehalem) (Up from 3.5x in [Haswell](https://en.wikichip.org/wiki/intel/microarchitectures/haswell))
* Mainstream chipset
  + [Lynx Point](https://en.wikichip.org/w/index.php?title=intel/chipsets/lynx_point&action=edit&redlink=1) → [Sunrise Point](https://en.wikichip.org/w/index.php?title=intel/chipsets/sunrise_point&action=edit&redlink=1)
  + Bus/Interface to Chipset
    - [DMI 3.0](https://en.wikichip.org/w/index.php?title=intel/direct_media_interface&action=edit&redlink=1) (from 2.0)
      * Skylake S and Skylake H cores, connected by 4-lane DMI 3.0
      * [Skylake Y](https://en.wikichip.org/wiki/intel/cores/skylake_y) and Skylake U cores have chipset in the same package (simplified [OPIO](https://en.wikichip.org/w/index.php?title=intel/on_package_i/o&action=edit&redlink=1))
      * Increase in transfer rate from 5.0 GT/s to 8.0 GT/s (~3.93GB/s up from 2GB/s) per lane
      * Limits motherboard trace design to 7 inches max from (down from 8) from the CPU to chipset
  + PCIe & DMI upgraded to Gen3
  + More I/O (configurable as PCIe/SATA/USB3)
  + Lower-power I/O (eMMC, UFS, SDXC)
  + CSI-2 for the integrated IPU (mobile SKUs)
  + Intel Sensor Solution Hub integrationLarger Line Fill Buffer?
* [System Agent](https://en.wikichip.org/w/index.php?title=System_Agent&action=edit&redlink=1)
  + New Image Processing Unit (IPU)
    - Incorporates an [image signal processor](https://en.wikichip.org/w/index.php?title=image_signal_processor&action=edit&redlink=1) (ISP)
    - Mobile client models only
  + Can now have its own variable voltage and frequency
* Core
  + Front End
    - Larger legacy pipeline delivery (5 µOPs, up from 4)
      * Another simple decoder has been added.
    - Allocation Queue (IDQ)
      * Larger delivery (6 µOPs, up from 4)
      * 2.28x larger buffer (64/thread, up from 56)
      * Partitioned for each active threads (from unified)
    - Improved [branch prediction unit](https://en.wikichip.org/w/index.php?title=branch_prediction_unit&action=edit&redlink=1)
      * reduced penalty for wrong direct jump target
      * No specifics were disclosed
    - µOP Cache
      * instruction window is now 64 Bytes (from 32)
      * 1.5x bandwidth (6 µOPs/cycle, up from 4)
  + Execution Engine
    - Larger [re-order buffer](https://en.wikichip.org/w/index.php?title=re-order_buffer&action=edit&redlink=1) (224 entries, up from 192)
    - Larger scheduler (97 entries, up from 64)
      * Larger Integer Register File (180 entries, up from 168)
      * Larger Retire (4 µOPs/cycle/thread, up from 4 µOPs/cycle/core)?
  + Memory Subsystem
    - Larger store buffer (56 entries, up from 42)
    - [L2$](https://en.wikichip.org/w/index.php?title=L2$&action=edit&redlink=1) was changed from 8-way to 4-way set associative
    - Page split load penalty reduced 20-fold
    - Larger Write-back buffer
* Memory
  + Support for faster DDR-2400 memory
  + [L3$](https://en.wikichip.org/w/index.php?title=L3$&action=edit&redlink=1) re-gained 512 KiB/core (See [§eDRAM architectural changes](https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(client)#eDRAM_architectural_changes) for the reason)
  + A new coherent [cache](https://en.wikichip.org/w/index.php?title=cache&action=edit&redlink=1) fabric implementation
    - doubles the throughput of the last level cache (LLC, L3$ in this case) miss handling
    - 50% improvement in bandwidth/watt
    - new [eDRAM](https://en.wikichip.org/w/index.php?title=eDRAM&action=edit&redlink=1" \o "eDRAM (page does not exist)) cache architecture for higher bandwidth
* TLBs
  + ITLB
    - 4 KiB page translations was changed from 4-way to 8-way associative
  + STLB
    - 4 KiB + 2 MiB page translations was changed from 6-way to 12-way associative
* Electrical
  + The fully integrated voltage regulator (FIVR) is moved back to the motherboard
    - Originally intended to be a cost-cutting measure by moving the FIVR on-die as well as making it more efficient, the move resulted in unintentionally making the FIVR the limiting factor when it came to overclocking.
  + DMI/PEG are now on a discrete clock domain with BCLK sitting on its own domain with full-range granularity (1 MHz intervals)
* Testability
  + New support for [Direct Connect Interface](https://en.wikichip.org/w/index.php?title=intel/direct_connect_interface&action=edit&redlink=1) (DCI), a new debugging transport protocol designed to allow debugging of closed cases (e.g. laptops, embedded) by accessing things such as [JTAG](https://en.wikichip.org/w/index.php?title=JTAG&action=edit&redlink=1" \o "JTAG (page does not exist))through any [USB 3](https://en.wikichip.org/w/index.php?title=USB_3&action=edit&redlink=1) port.
* [Gen 9 GPUs](https://en.wikichip.org/wiki/intel/microarchitectures/gen9)
  + Improved underlying implementation of the memory QoS for higher resolution displays and the integrated [image signal processor](https://en.wikichip.org/w/index.php?title=image_signal_processor&action=edit&redlink=1) (ISP)
    - Allow for higher concurrent bandwidth
  + Skylake retires VGA support, multi-monitor support for up to 3 displays via HDMI 1.4, DP 1.2, and eDP 1.3 interfaces.
  + Direct X 12, OpenCL 2.0, OpenGL 4.4
  + Up to 24 EUs GT2 (same as [Haswell](https://en.wikichip.org/wiki/intel/microarchitectures/haswell)); 48 EUs for GT3, and up to 72 EUs on [Iris Pro Graphics](https://en.wikichip.org/wiki/intel/iris_pro_graphics)
    - 1,152 GFLOPS @ 1 GHz

#### CPU changes[[edit](https://en.wikichip.org/w/index.php?title=intel/microarchitectures/skylake_(client)&action=edit&section=10" \o "Edit section: CPU changes)]

* Most ALU operations have 4 op/cycle 1 for 8 and 32-bit registers. 64-bit ops are still limited to 3 op/cycle. (16-bit throughput varies per op, can be 4, 3.5 or 2 op/cycle).
* MOVSX and MOVZX have 4 op/cycle throughput for 16->32 and 32->64 forms, in addition to Haswell's 8->32, 8->64 and 16->64 bit forms.
* ADC and SBB have throughput of 1 op/cycle, same as Haswell.
* Vector moves have throughput of 4 op/cycle (move elimination).
* Not only zeroing vector vpXORxx and vpSUBxx ops, but also vPCMPxxx on the same register, have throughput of 4 op/cycle.
* Vector ALU ops are often "standardized" to latency of 4. for example, vADDPS and vMULPS used to have L of 3 and 5, now both are 4.
* Fused multiply-add ops have latency of 4 and throughput of 0.5 op/cycle.
* Throughput of vADDps, vSUBps, vCMPps, vMAXps, their scalar and double analogs is increased to 2 op/cycle.
* Throughput of vPSLxx and vPSRxx with immediate (i.e. fixed vector shifts) is increased to 2 op/cycle.
* Throughput of vANDps, vANDNps, vORps, vXORps, their scalar and double analogs, vPADDx, vPSUBx is increased to 3 op/cycle.
* vDIVPD, vSQRTPD have approximately twice as good throughput: from 8 to 4 and from 28 to 12 cycles/op.
* Throughput of some MMX ALU ops (such as PAND mm1, mm2) is decreased to 2 or 1 op/cycle (users are expected to use wider SSE/AVX registers instead).

#### New instructions[[edit](https://en.wikichip.org/w/index.php?title=intel/microarchitectures/skylake_(client)&action=edit&section=11" \o "Edit section: New instructions)]

*See also:*[*Server Skylake's New instructions*](https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)#New_instructions)

Skylake introduced a number of [new instructions](https://en.wikichip.org/wiki/x86/extensions):

* [SGX1](https://en.wikichip.org/w/index.php?title=x86/sgx1&action=edit&redlink=1) - Software Guard Extensions, Version 1
* [MPX](https://en.wikichip.org/w/index.php?title=x86/mpx&action=edit&redlink=1) -Memory Protection Extensions
* [XSAVEC](https://en.wikichip.org/w/index.php?title=x86/xsavec&action=edit&redlink=1) - Save processor extended states with compaction to memory
* [XSAVES](https://en.wikichip.org/w/index.php?title=x86/xsaves&action=edit&redlink=1) - Save processor supervisor-mode extended states to memory.
* [CLFLUSHOPT](https://en.wikichip.org/w/index.php?title=x86/clflushopt&action=edit&redlink=1) - Flush & Invalidates memory operand and its associated cache line (All L1/L2/L3 etc..)